Optimal Score Estimation via Empirical Bayes Smoothing

Andre Wibisono, **Yihong Wu**, Kaylee Yingxi Yang

Yale University

ABSTRACT

We study the problem of estimating the score function of an unknown probability distribution $\rho^*$from n independent and identically distributed observations in d dimensions. Assuming that $\rho^*$is subgaussian and has a Lipschitz-continuous score function s^*, we establish the optimal rate of $\widetilde{\Theta}(n^{-\frac{2}{d+4}})$ for this estimation problem under the loss function $\| \hat{s} - s^* \|^{2}_{L2(\rho *)}$that is commonly used in the score matching literature, highlighting the curse of dimensionality where sample complexity for accurate score estimation grows exponentially with the dimension d. Leveraging key insights in empirical Bayes theory as well as a new convergence rate of smoothed empirical distribution in Hellinger distance, we show that a regularized score estimator based on a Gaussian kernel attains this rate, shown optimal by a matching minimax lower bound. We also discuss extensions to estimating β-Hölder continuous scores with $\beta \leq 1$, as well as the implication of our theory on the sample complexity of score-based generative models.

Keywords: Score estimation, kernel density estimation, empirical Bayes, Hellinger distance

Back to Sessions List

Provable Statistical and Computational Efficiency of Diffusion Models

Changxiao Cai, Gen Li

University of Michigan, Chinese University of Hong Kong

ABSTRACT

Score-based diffusion models have emerged as a foundational paradigm for modern generative modeling, achieving remarkable success across diverse applications from image synthesis to scientific computing. Despite their empirical prominence, fundamental questions about their theoretical foundations remain: How efficient can diffusion samplers be? What are the fundamental statistical limits of these samplers? In this talk, I will present recent theoretical advances that address both the computational and statistical frontiers of diffusion models. First, I will introduce a novel accelerated stochastic sampler that provably reduces iteration complexity under minimal assumptions, offering sampling speedups without sacrificing statistical optimality. Second, I will present the first comprehensive end-to-end analysis for deterministic ODE-based samplers, establishing (near-)minimax optimal statistical guarantees under mild assumptions on the target distribution. Together, these results provide a rigorous mathematical foundation that narrows the gap between the practical success and theoretical understanding of diffusion models. This is joint work with Gen Li.

Paper 1: https://arxiv.org/abs/2410.23285
Paper 2: https://arxiv.org/abs/2503.09583

Keywords: diffusion models, sampling, minimax optimiality, probability flow ODE, training-free acceleration

Back to Sessions List

Transformers Provably Learn Chain-of-Thought Reasoning with Length Generalization

Yuejie Chi

Department of Statistics and Data Science, Yale University

ABSTRACT

The ability to reason lies at the core of artificial intelligence (AI), and challenging problems usually call for deeper and longer reasoning to tackle. A crucial question about AI reasoning is whether models can extrapolate learned reasoning patterns to solve harder tasks with longer chain-of-thought (CoT). In this work, we present a theoretical analysis of transformers learning on synthetic state-tracking tasks with gradient descent. We mathematically prove how the algebraic structure of state-tracking problems governs the degree of extrapolation of the learned CoT. Specifically, our theory characterizes the length generalization of transformers through the mechanism of attention concentration, linking the retrieval robustness of the attention layer to the state-tracking task structure of long-context reasoning. Moreover, for transformers with limited reasoning length, we prove that a recursive self-training scheme can progressively extend the range of solvable problem lengths. To our knowledge, we provide the first optimization guarantee that constant-depth transformers provably learn -complete problems with CoT, significantly going beyond prior art confined in , unless the widely held conjecture fails. Finally, we present a broad set of experiments supporting our theoretical results, confirming the length generalization behaviors and the mechanism of attention concentration.

Keywords: transformers, chain-of-thought, length generalization

Back to Sessions List